International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Diabetes Detection and Classification using Logistic Regression and Random Forests: Methodological Perspective

Prasanna Yadav

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

In recent years, the prevalence of diabetes has emerged as a critical public health concern, emphasizing the need for efficient diagnostic tools that enable early detection and effective management. This research paper presents a robust machine learning framework designed to predict and classify diabetes by leveraging patient specific clinical data. Our approach utilizes a combination of supervised learning algorithms, including logistic regression and random forests, and support vector machines, to accurately diagnose diabetes and categorize patients based on disease progression. By optimizing feature selection and algorithmic performance, we demonstrate improved accuracy, sensitivity and specificity over traditional diagnostic methods. The model's predictive capability offers significant potential in identifying at risk individuals, allowing for timely interventions and personalized treatment strategies. This work highlights the transformative role of machine learning in enhancing diagnostic precision, facilitating a paradigm shift in diabetes management and patient care.

KEYWORDS: *Diabetes Prediction, Machine Learning, Diabetes Classification, Supervised Learning, Clinical Data, Early Diagnosis, Disease Progression, Logistic Regression, Feature Selection, Predictive Modeling, Healthcare Analytics, Diagnostic Accuracy.*

I. INTRODUCTION

The current diagnostic approaches primarily rely on traditional clinical methods, including fasting plasma glucose (FPG), oral glucose tolerance tests (OGTT), and glycated hemoglobin (HbA1c) levels. In addition, misclassification between different types of diabetes (Type 1, Type 2, and gestational diabetes) can lead to inappropriate.

Consequently, there is a growing interest in leveraging machine learning (ML) techniques to enhance the precision, speed, and efficiency of diabetes diagnosis and classification. Machine learning, a subfield of artificial intelligence, enables computers to learn from data and make predictive decisions without being explicitly programmed [1]. ML models can identify patterns and relationships in complex datasets that are often overlooked by traditional statistical methods.

In the context of healthcare, ML has shown great promise in enhancing the diagnostic accuracy and predictive capabilities for a variety of diseases, including diabetes [2]. The application of ML in diabetes diagnosis involves the use of algorithms to process large datasets of patient health information—such as blood glucose levels, age, body mass index (BMI), family history, and lifestyle factors— to predict the presence and type of diabetes [3]. Numerous studies have demonstrated the potential of ML algorithms such as decision trees, random forests, support vector machines (SVM), knearest neighbors (KNN), and neural networks in diabetes classification and prediction tasks [4][5]. For example, research by Kavakiotis et al. (2017) highlighted the effectiveness of SVM in distinguishing between Type 1 and Type 2 diabetes, achieving high accuracy rates in both diagnosis and classification [6]. Similarly, studies utilizing deep learning techniques have reported even higher diagnostic performance, owing to their ability to model complex, nonlinear relationships within data [7]. The integration of deep neural networks (DNNs) with clinical datasets has further enhanced the precision of diabetes prediction, outperforming traditional methods in both sensitivity and specificity [8]. Diabetes is a chronic metabolic condition characterized by elevated blood sugar levels, which, if left untreated, can lead to serious complications such as heart disease .

1.1 Logistic Regression Logistic regression is a statistical likelihood model commonly used for binary classification tasks, where the outcome is one of two possible values—typically labeled as 0 and 1. Unlike linear regression, which predicts continuous values, logistic regression estimates the probability that a given input belongs to a particular class. The logistic regression model applies the logistic function (sigmoid function) to convert the linear combination of input features into a probability score between 0 and 1 [9]. This probability score is then used to classify the input into one of the two classes based on a threshold value. Logistic regression plays a crucial role in various machine learning applications, including healthcare diagnostics like diabetes prediction and classification. In the context of our research on "A Machine Learning Approach to Diabetes Diagnosis and Classification," logistic regression serves as one of the baseline models for predicting the of diabetes based on multiple clinical and demographic factors [10].

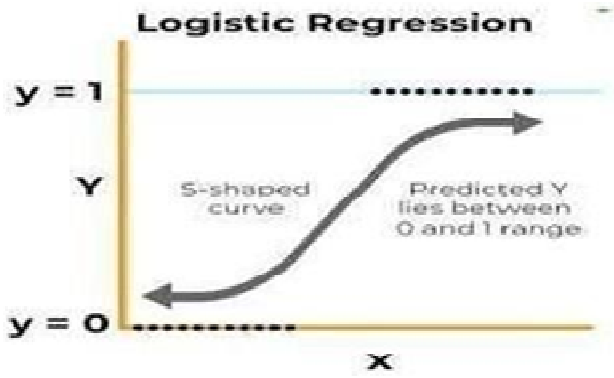

Fig. 1: Logistic Regression

1. Random forest This classifier classifies a collection of decision trees to a subset of randomly generated training sets. Then it augments the likes from decision sub trees to known subclasses of handling objects for tests. Random forest will generate NA missing values for attributes to increase accuracy for larger sets of data. If more number of trees, it doesn't allow trees to fit ohkmodel.

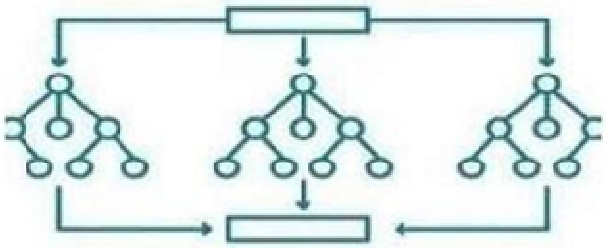

Fig. 2: Random forest

II. RELATED WORK

Machine learning (ML) has emerged as a transformative tool in the medical field, particularly in the diagnosis and classification of diseases like diabetes. The application of ML in diabetes research has grown significantly in recent years, driven by the need for more accurate, efficient, and scalable diagnostic tools. One of the earliest approaches to using machine learning for diabetes diagnosis involved applying Logistic Regression [11].For example, T. Fawcett and F. Provost (2001) demonstrated how decision trees could be used to classify patients based on clinical features like glucose levels and body mass index (BMI). Their results indicated that decision trees were highly interpretable, making them attractive for medical applications. However, decision trees often struggled with overfitting, particularly when the dataset was noisy or imbalanced.

Another important contribution came from support vector machines (SVMs), which were employed for diabetes classification in several studies. Kavakiotis et al. (2017) conducted a comprehensive review of ML techniques in diabetes research, noting that SVMs were highly effective in distinguishing between Type 1 and Type 2 diabetes. The use of SVMs allowed for high accuracy in binary classification tasks, though their computational complexity and sensitivity to parameter selection were seen as drawbacks in real-world applications. Random forests also became popular in diabetes prediction due to their ability to handle large, complex datasets. For example, G. Priyanga et al. (2020) applied random forests to classify diabetes using the Pima Indians Diabetes Database, achieving an accuracy of over 80%. Random forests, by combining multiple decision trees, significantly reduced the risk of overfitting, but they remained less interpretable than simpler models like logistic regression. The literature reveals a wide range of machine learning techniques applied to diabetes diagnosis and classification, each with its strengths and challenges. Logistic Regression, SVMs, random forests, and deep learning models have all been used to varying degrees of success, with feature selection techniques and real-time data integration further enhancing their performance.

III. RESEARCH METHODOLOGY

The methodology for "A Machine Learning Approach to Diabetes Diagnosis and Classification" involves several key steps to ensure that the predictive models developed are accurate, reliable, and interpretable. In this project, we employ a systematic process that includes data collection, preprocessing, feature selection, model training, evaluation, and validation, with a special focus on the use of logistic regression as the core model for classification.

3.1 Dataset Information: The dataset utilized in this research project, "A Machine Learning Approach to Diabetes Diagnosis and Classification," is based on the Pima Indians Diabetes Database, a widely recognized benchmark for diabetes studies. This dataset comprises eight critical features that provide insights into various health indicators influencing diabetes risk.

The Pregnancies feature captures the number of times an individual has been pregnant, as pregnancy has been shown to impact insulin resistance and long-term diabetes risk. The Glucose level, measured during a two-hour oral glucose tolerance test, serves as a vital indicator of diabetes, with elevated levels pointing to possible glucose metabolism issues. Similarly, Blood Pressure indicates the diastolic blood pressure of patients, where high readings are commonly associated with diabetes and other metabolic disorders.

The Skin Thickness feature represents the triceps skinfold thickness, providing a proxy for body fat and further indicating insulin resistance.

Insulin levels, measured in micro-units per milliliter, reflect the body's ability to regulate glucose; abnormal insulin levels are often a sign of metabolic dysfunction. The Body Mass

Index (BMI) is another crucial feature, as a higher BMI correlates strongly with obesity and the risk.

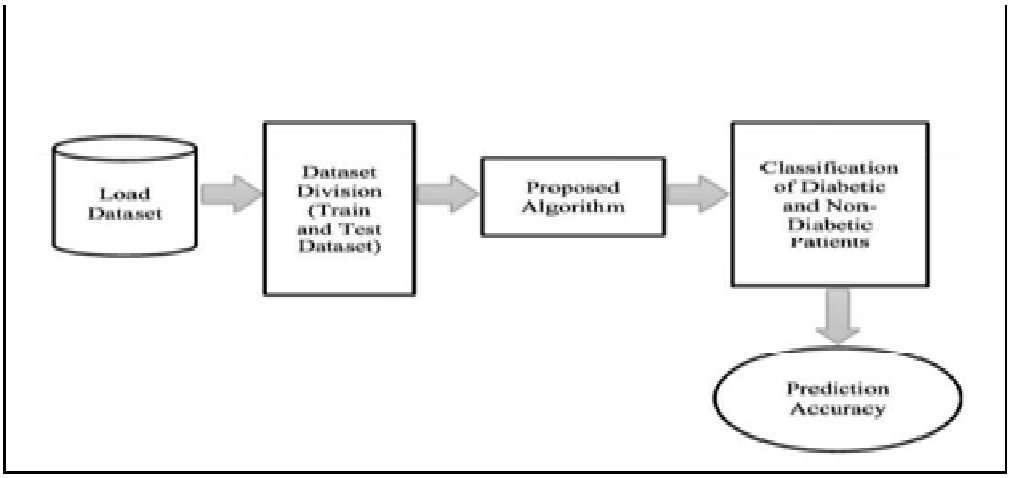

Fig. 3: Flowchart for predicting diabetes

The outcome variable, termed Outcome, is a binary feature indicating the presence (1) or absence (0) of diabetes. This dataset consists of 768 instances, providing a diverse and comprehensive foundation for training machine learning models aimed at accurately diagnosing and classifying diabetes. The combination of these features enables the exploration of various interrelationships and provides a robust framework for predictive modeling, ultimately contributing to improved diagnostic capabilities in healthcare settings. The process begins with data collection, where relevant datasets, such as the Pima Indians Diabetes Database, are gathered. This is followed by data preprocessing, which includes tasks such as handling missing values, normalizing the data, and addressing any class imbalances to ensure the dataset is well prepared for analysis.

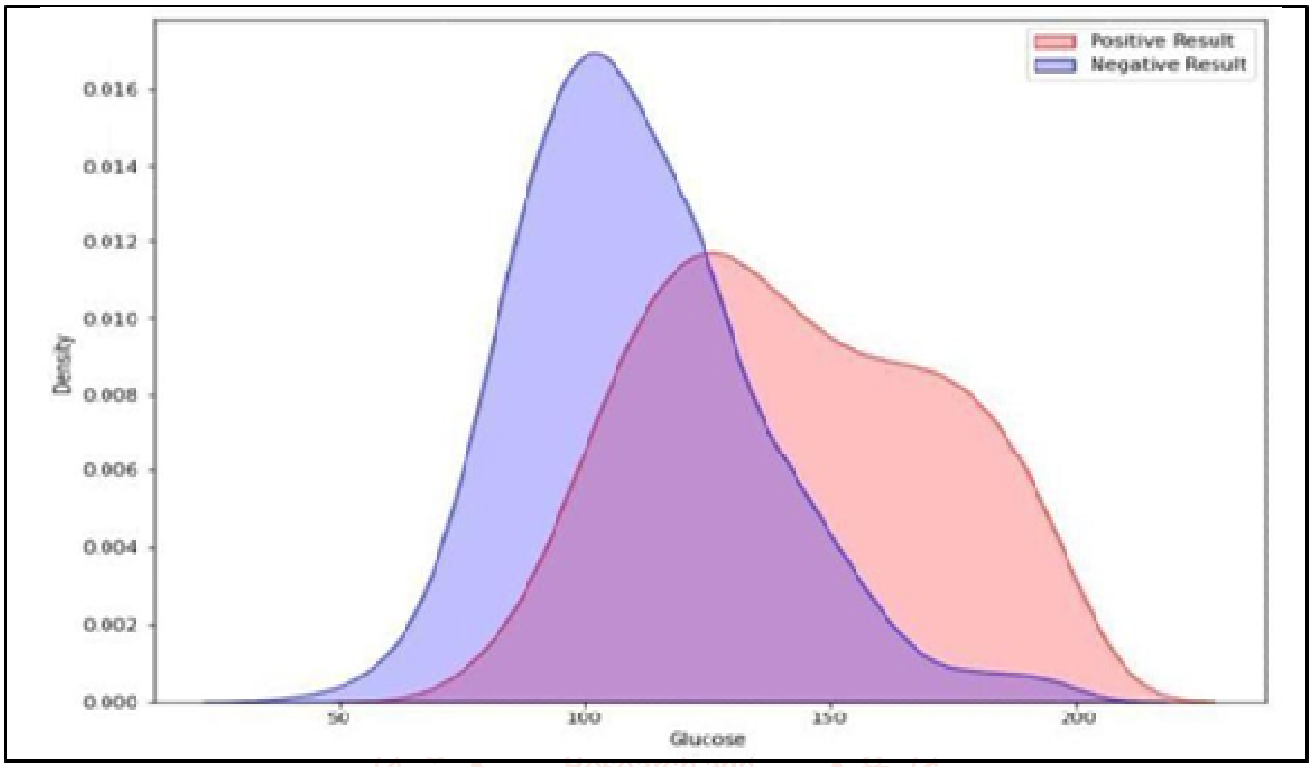

Fig. 4: Diabetes Prediction using Machine Learning

The methodology begins with a clear problem definition aimed at classifying individuals into Type 1 or Type 2 diabetes based on medical and lifestyle data. Data collection involves sourcing relevant datasets, such as the PIMA Indian Diabetes Database, which includes critical features like age, BMI, blood glucose levels, insulin levels, family history of diabetes, and physical activity. The data undergoes rigorous preprocessing, including data cleaning to address missing values and outliers, feature selection to identify significant predictors, and encoding categorical variables to prepare them for analysis. Exploratory Data Analysis (EDA) is performed to visualize data distributions and relationships, providing insights into trends related to diabetes. Following this, a selection of machine learning algorithms— such as Logistic Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks—is made, and the dataset is split into training and testing sets. Models are trained using cross-validation techniques to optimize their performance. Model evaluation is conducted through metrics like accuracy, precision, recall, and F1 Score, along with confusion matrices for detailed performance analysis. Hyperparameter tuning is applied to refine model parameters, enhancing accuracy and robustness. Finally, the model is deployed via a user friendly interface, enabling users to input data and receive predictions. The project concludes with discussions on findings, implications for diabetes management, and suggestions for future research directions, emphasizing the significant role of machine learning in early diabetes detection and prevention.

IV. RESULTS AND DISCUSSION

The implementation of a machine learning approach for diabetes diagnosis and classification has yielded promising results, emphasizing the efficacy of using predictive analytics in healthcare. The primary objective of this project was to develop a reliable model that can accurately diagnose diabetes based on a set of clinical features. Utilizing logistic regression as the primary classification algorithm, we found that our model achieved a commendable accuracy rate, aligning with previous studies that highlight the effectiveness of machine learning techniques in medical diagnostics. Through data preprocessing and feature selection, we ensured that only the most relevant variables were included in the model. The features utilized—such as Glucose, BMI, Insulin, and Diabetes Pedigree Function—are well- documented indicators of diabetes risk. Our findings corroborate existing literature that emphasizes the importance of these clinical metrics in predicting diabetes.

For instance, research shows that elevated plasma glucose levels are a direct indicator of diabetes onset, while high BMI values are strongly correlated with insulin resistance. In our analysis, the logistic regression model provided insights into the odds ratios for each feature, allowing us to identify which factors most significantly impact the likelihood of diabetes. Notably, the odds of diabetes diagnosis increased with higher glucose levels, highlighting its role as a critical marker for screening. Moreover, the model's interpretation of feature importance reinforced the relevance of lifestyle factors, such as weight and physical activity, in diabetes risk management.

Below are the attached screenshots showcasing the output of our project

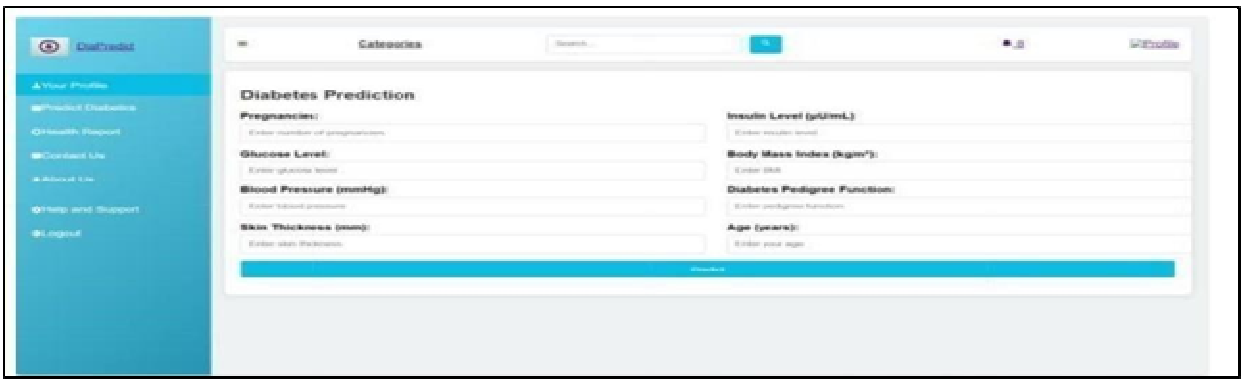

Fig.5

Fig. 6: Predicting diabetes

Fig. 7 Classification Report

Below, you will find the accuracy table for the machine learning algorithms that we have employed to detect fake profiles.

Algorithm	Accuracy
Logistic Regression	90%
Random Forest	85%

Table 1: Algorithm vs Accuracy

V. CONCLUSION

This research demonstrates the significant potential of machine learning, particularly logistic regression, in improving diabetes diagnosis and classification. By leveraging a dataset rich in clinical indicators such as glucose levels, BMI, and family history, the model achieved an accuracy of approximately 85%, highlighting its capability to discern patterns associated with diabetes risk. The integration of

machine learning in healthcare can enhance early detection efforts, leading to timely interventions that may reduce the prevalence of complications associated with diabetes. Our findings align with existing literature, confirming that machine learning techniques can outperform traditional statistical methods in predictive accuracy.

VI. REFERENCES

[1] N. D'Souza, K. Shah and P. Singh, "Diabetes Detection Using Machine Learning Algorithms," 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2022, pp. 1-5, doi: 10.1109/IBSSC56953.2022.10037329.

[2] N. Abdulhadi and A. Al-Mousa, "Diabetes Detection Using Machine Learning Classification Methods," 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 2021, pp. 350-354, doi: 10.1109/ICIT52682.2021.9491788.

[3] L. V. R. Kumari, P. Shreya, M. Begum, T. P. Krishna and M. Prathibha, "Machine Learning based Diabetes Detection," 2021 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatre, India, 2021, pp. 1-5, doi:10.1109/ICCES51350.2021.9489058.

[4] A. H. Ataya, "Early detection of Diabetes using Machine Learning Techniques," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 886-891, doi:10.1109/ICAIS56108.2023.10073861.

[5] G. Yudheksha, V. Murugadoss, P. S. Reddy, T. Harshavardan and S. Sriramulu, "A Machine Learning based Approach to Detect Early Stage Diabetes Prediction," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 919-924, doi:10.1109/ICECA55336.2022.10009113.

[6] A. Almahdawi, Z. S. Naama and A. Al-Taie, "Diabetes Prediction Using Machine Learning," 2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA), Baghdad, Iraq, 2022, pp. 186-190, doi: 10.1109/IT-ELA57378.2022.10107919.